Cloned DNA sequences related to the genomic RNA of lymphadenopathy-associated-virus (LAV) and proteins encoded by said LAV genomic RNA

The invention relates to cloned DNA sequences indistinguishable from genomic RNA and DNA of lymphadenopathy-associated virus (LAV), a process for their preparation and their uses. It relates more particularly to stable probes including a DNA sequence which can be used for the detection of the LAV virus or related viruses or DNA proviruses in any medium, particularly biological samples containing any of them. The invention also relates to polypeptides, whether glycosylated or not, encoded by said DNA sequences.

Lymphadenopathy-associated virus (LAV) is a human retrovirus first isolated from the lymph node of a homosexual patient with lymphadenopathy syndrome, frequently a prodrome or a benign form of acquired immune deficiency syndrome (AIDS). Subsequently, other LAV isolates were recovered from patients with AIDS or pre-AIDS. All available data are consistent with the virus being the causative agent of AIDS.

A method for cloning such DNA sequences has already been disclosed in British Patent Application Nr. 84 23659, filed on September 19, 1984. Reference is hereafter made to that application as concerns subject matter in common with the further improvements to the invention disclosed herein.

The present invention aims at providing additional new means which should not only also be useful for the detection of LAV or related viruses, (hereafter more generally referred to as "LAV viruses"), but also have more versatility, particularly in detecting specific parts of the genomic RNA of said viruses whose expression products are not always directly detectable by immunological methods.

The present invention further aims at providing

polypeptides containing sequences in common with polypeptides encoded by the LAV genomic RNA. It relates even more particularly to polypeptides comprising antigenic determinants included in the proteins encoded and expressed by the LAV genome occurring in nature. An additional object of the invention is to further provide means for the detection of proteins related to LAV virus, particularly for the diagnosis of AIDS or pre-AIDS or, to the contrary, for the detection of antibodies against the LAV virus or proteins related therewith, particularly in patients afflicted with AIDS or pre-AIDS or more generally in asymtomatic carriers and in blood-related products. Finally, the invention also aims at providing immunogenic polypeptides, and more particularly protective polypeptides for use in the preparation of vaccine compositions against AIDS or related syndromes.

The present invention relates to additional DNA fragments, hybridizable with the genomic RNA of LAV as they will be disclosed hereafter, as well as with additional cDNA variants corresponding to the whole genomes of LAV viruses. It further relates to DNA recombinants containing said DNAs or cDNA fragments.

The invention relates more particularly to a cDNA variant corresponding to the whole of LAV retroviral genomes, which is characterized by a series of restriction sites in the order hereafter (from the 5' end to the 3' end).

The coordinates of the successive sites of the whole LAV genome (restriction map) are indicated hereafter too, with respect to the Hind III site (selected as of coordinate 1) which is located in the R region. The coordinates are estimated with an accuracy of ± 200 bp :

Hind III	0
Sac I	50
Hind III	520
Pst I	800
Hind III	1 100

Bgl II	1 500
Kpn I	3 500
Kpn I	3 900
Eco RI	4 100
Eco RI	5 300
Sal I	5 500
Kpn I	6 100
Bgl II	6 500
Bgl II	7 600
Hind III	7 850
Bam HI	8 150
Xho I	8 600
Kpn I	8 700
Bgl II	8 750
Bgl II	9 150
Sac I	9 200
Hind III	9 250

Another DNA variant according to this invention optionally contains an additional Hind III approximately at the 5 550 coordinate.

Reference is further made to fig. 1 which shows a more detailed restriction map of said whole DNA (λJ19).

An even more detailed nucleotide sequence of a preferred DNA according to the invention is shown in Figs. 4-12 hereafter.

The invention further relates to other preferred DNA fragments which will be referred to hereafter.

Additional features of the invention will appear in the course of the non-limitative disclosure of additional features of preferred DNAs of the invention, as well as of preferred polypeptides according to the invention. Reference will further be had to the drawings in which :

- fig. 1 is the restriction map of a complete LAV genome (clone λJ19) ;
- figs. 2 and 3 show diagrammatically parts of the three

possible reading phases of LAV genomic RNA, including the open reading frames (ORF) apparent in each of said reading phases ;

- figs. 4-12 show the successive nucleotide ~~nucleotidic~~ sequences of a complete LAV genome. The possible peptide ~~peptidic~~ sequences in relation to the three possible reading phases related to the nucleotide ~~nucleotidic~~ sequences shown are also indicated ;
- figs. 13-18 reiterate the sequence of part of the LAV genome containing the genes coding for the envelope ~~envelope~~ proteins, with particular boxed peptide ~~peptidic~~ sequences which correspond ~~responds~~ to groups which normally carry glycosyl groups.

The sequencing and determination of sites of particular interest were ~~was~~ carried out on a phage recombinant corresponding to λJ19 disclosed in the abovesaid British Patent application Nr. 84 23659. A method for preparing it is disclosed in that application.

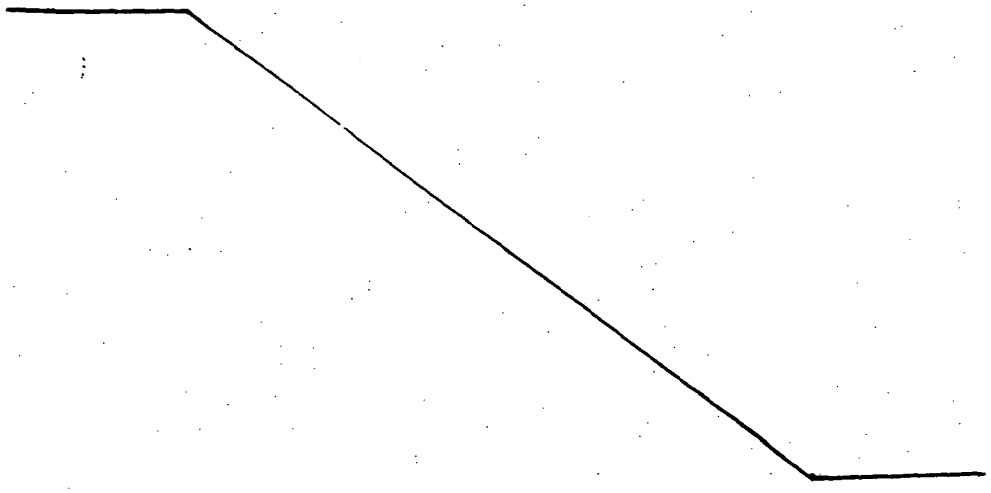
The whole recombinant phage DNA of clone λJ19 (disclosed in the earlier application) was sonicated according to the protocol of DEININGER (1983), Analytical Biochem. 129, 216. The ~~the~~ DNA was repaired by a Klenow reaction for 12 hours at 16°C. The DNA was electrophoresed through 0.8 % agarose gel and DNA in the size range of 300-600 bp was cut out and electroeluted and precipitated. Resuspended DNA (in 10 mM Tris, pH 8 ; 0,1 mM EDTA) was ligated into M13mp8 RF DNA (cut by the restriction enzyme SmaI ~~SmaI~~ and subsequently alkaline phosphated), using T4 DNA- and RNA-ligases (Maniatis T et al (1982) - Molecular cloning - Cold Spring Harbor Laboratory). An *E. coli* strain designated as TG1 was used for further study. This strain has the following genotype :

Δlac pro, supE, thi.F'traD36, proAB, $lacI^q$, ZΔM15,r^-

This *E. coli* TGI strain has the peculiarity of enabling recombinants to be recognized easily. The blue colour of the cells transfected with plasmids which did

not recombine with a fragment of LAV DNA is not modified. To the contrary cells transfected by a recombinant plasmid containing a LAV DNA fragment yield white colonies. The technique which was used is disclosed in Gene (1983), 26, 101.

This strain was transformed with the ligation mix using the Hanahan method (Hanahan D (1983) J. Mol. Biol. 166, 557). Cells were plated out on tryptone-agarose plate with IPTG and X-gal in soft agarose. White plaques were either picked and screened or screened directly in situ using nitrocellulose filters. Their DNAs were hybridized with nick-translated DNA inserts of pUC18 HindIII ~~Hind III~~ subclones of λJ19. This ~~this~~ permitted the isolation of the plasmids or subclones of λ which are identified in the table hereafter. In relation to this table it should also be noted that the designation of each plasmid is followed by the deposition number of a cell culture of E. coli TGI containing the corresponding plasmid at the "Collection Nationale des Cultures de Micro-organismes" (C.N.C.M.) of the Pasteur Institute in Paris, France. A non-transformed TGI cell line was also deposited at the C.N.C.M. under Nr. I-364. All these deposits took place on November 15, 1984. The sizes of the corresponding inserts derived from the LAV genome have also been indicated.



TABLE

Essential features of the recombinant plasmids

- pJ19 - 1 plasmid (I-365) 0.5 kb

Hind III - Sac I - Hind III

- pJ19 - 17 plasmid (I-367) 0.6 kb

Hind III - Pst 1 - Hind III

- pJ19 - 6 plasmid (I-366) 1.5 kb

Hind III (5')
Bam HI
Xho I
Kpn I
Bgl II
Sac I (3')
Hind III

- pJ19-13 plasmid (I-368) 6.7 kb

Hind III (5')
Bgl II
Kpn I
Kpn I
Eco RI
Eco RI
Sal I
Kpn I
Bgl II
Bgl II
Hind III (3')

Positively hybridizing M13 phage plates were grown up for 5 hours and the single-stranded DNAs were extracted.

M13mp8 subclones of λJ19 DNAs were sequenced according to the dideoxy method and technology devised by Sanger et al. Sanger et al (1977), Proc. Natl. Acad. Sci. USA, 74, 5463 and M13 cloning and sequencing handbook, AMERSHAM (1983). The 17-mer oligonucleotide primer α-^{35}SdATP (400Ci/mmol, AMERSHAM), and 0.5X-5X buffer gradient gels (Biggen M.D. et al (1983) Proc. Natl. Acad. Sci. USA, 50, 3963) were used. Gels were read and put into the computer under the programs of Staden (Staden R. (1982), Nucl. Acids Res. 10, 4731). All the appropriate references and methods can be found in the AMERSHAM M13 cloning and sequencing handbook.

The complete sequence of λJ19 was deduced from the experiments as further disclosed hereafter.

Figs. 4-12 provide the DNA nucleotide sequence of the complete genome of LAV. The numbering of the nucleotides starts from a left most HindIII restriction site (5' AAG...) of the restriction map. The numbering occurs in tens whereby the last zero number of each of the numbers occuring on the drawings is located just below the nucleotide corresponding to the nucleotides designated. That is, the nucleotide at position 10 is T, the nucleotide at position 20 is C, etc..

Above each of the lines of the successive nucleotide sequences there are provided three lines of single letters corresponding to the amino acid sequence deduced from the DNA sequence (using the genetic code) for each of the three reading phases, whereby said single letters have the following meanings.

A : alanine
R : arginine
K : lysine
H : histidine
C : cysteine

M : méthionine
W : tryptophan
F : phenylalanine
Y : tyrosine
L : leucine
V : valine
I : isoleucine
G : glycine
T : thréonine
S : serine
E : glutamic acid
D : Aspartic acid
N : asparagine
Q : glutamine
P : proline.

The asterik signs "*" correspond to stop codons (i.e. TAA, TAG and TGA).

Starting above the first line of the DNA nucleotide ~~nucleotidic~~ sequence of fig. 4, the three reading phases are respectively marked "1", "2", "3", on the left hand side ~~handside~~ of the drawing. The same relative presentation of the three theoretical ~~theoritical~~ reading phases is then used over all ~~all over~~ the successive ~~successives~~ lines of the LAV nucleotide ~~nucleotidic~~ sequence.

Figs. 2 and 3 provide a diagrammatized representation of the lengths of the successive open reading frames corresponding to the successive reading phases (also referred to by numbers "1", "2" and "3" appearing in the left hand side ~~handside~~ part of fig. 2). The relative positions of these open reading frames (ORF) with respect to the nucleotide ~~nucleotidic~~ structure of the LAV genome is referred to by the scale of numbers representative of the respective positions of the corresponding nucleotides in the DNA sequence. The vertical bars correspond to the positions of the corresponding stop codons.

1) <u>The "gag gene" (or ORF-gag)</u>

The "gag gene" codes for core proteins.

Particularly it appears that a genomic fragment (ORF-gag) thought to code for the core antigens including the p25, p18 and p13 proteins is located between nucleotide ~~nucleotidic~~ position 236 (starting with 5' CTA GCG GAG 3') and nucleotide ~~nucleotidic~~ position 1759 (ending by CTCG TCA CAA 3'). The structure of the peptides or proteins encoded by parts of said ORF is deemed to be that corresponding to phase 2.

The methionine amino acid ~~aminoacid~~ "M" coded by the ATG at position 260-262 is the probable initiation methionine of the gag protein precursor. The end of ORF-gag and accordingly of gag protein appears to be located at position 1759.

The beginning of p25 protein, thought to start by a P-I-V-Q-N-I-Q-G-Q-M-V-H .... amino acid ~~aminoacid~~ sequence is thought to be coded for by the nucleotide ~~nucleotidic~~ sequence CCTATA..., starting at position 656.

Hydrophilic peptides in the gag open reading frame are identified hereafter. They are defined starting from amino acid ~~aminoacid~~ 1 = Met (M) coded by the ATG starting from 260-2 in the LAV ONA sequence.

Those hydrophilic peptides are

12-32	amino acids ~~aminoacids~~	inclusive
37-46	"	"
49-79	"	"
88-153	"	"
158-165	"	"
178-188	"	"
200-220	"	"
226-234	"	"
239-264	"	"
288-331	"	"
352-361	"	"
377-390	"	"
399-432	"	"
437-484	"	"
492-498	"	"

The invention also relates to any combination of these peptides.

2) The "pol gene" (or ORF-pol)

Figs. 4-12 also show that the DNA fragments extending from nucleotide ~~nucleotidic~~ position 1555 (starting with 5'TTT ~~TTT~~ ....3' to nucleotide ~~nucleotidic~~ position 5086 is thought to correspond to the pol gene. The polypeptidic structure of the corresponding polypeptides is deemed to be that corresponding to phase 1. It stops at position 4563 (end by 5'G GAT GAG GAT 3').

These genes are thought to code for the virus polymerase or reverse transcriptase.

3) The envelope gene (or ORF-env)

The DNA sequence thought to code for envelope proteins is thought to extend from nucleotide ~~nucleotidic~~ position 5670 (starting with 5'AAA GAG GAG A....3') up to nucleo~~tidic~~ide position 8132 (ending by ....A ACT AAA GAA 3'). Polypeptide ~~Polypeptidic~~ structures of sequences of the envelope protein correspond to those read according to the "phase 3" reading phase.

The start of env transcription is thought to be at the level of the ~~th~~ ATG codon at positions 5691-5693.

Additional features ~~feature~~ of the envelope protein coded by the env genes appear on figs. 13-18. These are to be considered as paired figs. 13 and 14 ; 15 and 16 ; 17 and 18, respectively.

It is to be mentioned that because of format difficulties,

Fig. 14 overlaps to some extent with fig. 13,

Fig. 16 overlaps to some extent with fig. 15,

Fig. 18 overlaps to some extent with fig. 17.

Thus, for instance, figs. 13 and 14 must be considered together. Particularly the sequence shown on the first line on the top of fig. 13 overlaps with the sequence shown on the first line on the top of fig. 14. In other words, the starting of the reading of the successive

sequences of the env gene as represented in figs. 13-18 involves first reading the first line at the top of fig. 13 then proceeding further with the first line of fig. 14. One then returns to the beginning of the second line of fig. 13, then again further proceeds with the reading of the second line of page 14, etc. The same observations then apply to the reading of the paired figs. 15 and 16, and paired figs. 17 and 18, respectively.

The locations of neutralizing epitopes are further apparent in figs. 13-18. Reference is more particularly made to the boxed groups of three letters included in the amino acid sequences of the envelope proteins (reading phase 3) which can be designated generally by the formula N-X-S or N-X-T, wherein X is any other possible amino acid. Thus, the initial protein product of the env gene is a glycoprotein of molecular weight in excess of 91,000. These groups are deemed to generally carry glycosylated groups. These N-X-S and N-X-T groups with attached glycosylated groups form together hydrophilic regions of the protein and are deemed to be located at the periphery of and to be exposed outwardly with respect to the normal conformation of the proteins. Consequently, they are considered as being epitopes which can efficiently be brought into play in vaccine compositions.

The invention thus concerns with more particularity peptide sequences included in the env proteins and excizable therefrom (or having the same amino acid structure), having sizes not exceeding 200 amino acids.

Preferred peptides of this invention (referred to hereafter as a, b, c, d, e, f) are deemed to correspond to those encoded by the nucleotide sequences which extend, respectively, between the following positions :

a) from about 6095 to about 6200
b) " " 6260 " " 6310
c) " " 6390 " " 6440
d) " " 6485 " " 6620

e) " " 6860 " " 6930

f) " " 7535 " " 7630

Other hydrophilic peptides in the env open reading frame are identified hereafter. ~~they~~ They are defined starting from ~~aminoacid~~ amino acid 1 = lysine (K) coded by the AAA at position 5670-2 in the LAV DNA sequence.

These hydrophilic peptides are

8-23 ~~aminoacids~~ amino acids inclusive

63-78 " "

82-90 " "

97-123 " "

127-183 " "

197-201 " "

239-296 " "

300-327 " "

334-381 " "

397-424 " "

466-500 " "

510-523 " "

551-577 " "

594-603 " "

621-630 " "

657-679 " "

719-758 " "

780-803 " "

The invention also relates to any combination of these peptides.

4) <u>The other ORF</u>

The invention further concerns DNA sequences which provide open reading frames defined as ORF-Q, ORF-R and as "1", "2", "3", "4", "5", the relative position of which appears more particularly in figs. 2 and 3.

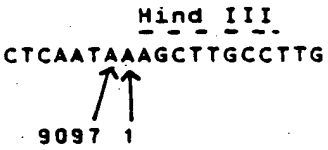
These ORFs have the following locations :

ORF-Q phase 1 start 4478 stop 5086

ORF-R " 2 " 8249 " 8896

ORF-1	"	1	"	5029	"	5316
ORF-2	"	2	"	5273	"	5515
ORF-3	"	1	"	5383	"	5616
ORF-4	"	2	"	5519	"	5773
ORF-5	"	1	"	7966	"	8279

The LTR (long terminal repeats) can be defined as lying between position 8560 and position 160 (end extending over position 9097/1). As a matter of fact the end of the genome is at 9097 and, because of the LTR structure of the retrovirus, links up with the beginning of the sequence :



The invention concerns more particularly all the DNA fragments which have been more specifically referred to hereabove and which correspond to open reading frames. It will be understood that the man skilled in the art will be able to obtain them all, for instance by cleaving an entire DNA corresponding to the complete genome of a LAV species, such as by cleavage by a partial or complete digestion thereof with a suitable restriction enzyme and by the subsequent recovery of the relevant fragments. The different DNAs disclosed in the earlier mentioned British Application can be resorted to also as a source of suitable fragments. The techniques disclosed hereabove for the isolation of the fragments which were then included in the plasmids referred to hereabove and which were then used for the DNA sequencing can be used.

Of course other methods can be used. Some of them have been ~~exemplified~~ exemplified in the earlier British Application. ~~reference is for instance~~ Reference is, for instance, made to the following methods.

a) DNA can be transfected into mammalian cells with appropriate selection markers by a variety of techniques, such as calcium phosphate precipitation, polyethylene

glycol, protoplast-fusion, etc..

b) DNA fragments corresponding to genes can be cloned into expression vectors for E. coli, yeast or mammalian cells and the resultant proteins purified.

c) The provival DNA can be "shot-gunned" (fragmented) into procaryotic expression vectors to generate fusion polypeptides. ~~Recombinant~~ Recombinants producing antigenically competent fusion proteins can be identified by simply screening the recombinants with antibodies against LAV antigens .

The invention also relates more specifically to cloned probes which can be made starting from any DNA fragment according to this invention, thus to recombinant DNAs containing such fragments, particularly any plasmids amplifiable in procaryotic or eucaryotic cells and carrying said fragments.

Using the cloned DNA fragments as a molecular hybridization probe - either by marking with radionucleotides or with fluorescent reagents - LAV virion RNA may be detected directly in the blood, body fluids and blood products (e.g. of the ~~antihemophylic~~ antihemophilic factors, such as Factor VIII concentrates) and vaccines, i.e. hepatitis B vaccine. It has already been shown that whole virus can be detected in culture supernatants of LAV producing cells. A suitable method for achieving that detection comprises immobilizing virus onto ~~said~~ a support, e.g. nitrocellulose filters, etc., disrupting the virion, and hybridizing with labelled (radiolabelled or "cold" fluorescent- or enzyme-labelled) probes. Such an approach has already been developed for Hepatitis B virus in peripheral blood (according to SCOTTO J. et al. Hepatology (1983), 3, 379-384).

Probes according to the invention can also be used for rapid screening of genomic DNA derived from the tissue of patients with LAV related symptoms to see if the proviral DNA or RNA is present in host tissue and other

tissues.

A method which can be used for such screening ~~comprise~~ Comprises the following steps : extraction of DNA from tissue, restriction enzyme cleavage of said DNA, electrophoresis of the fragments and Southern blotting of genomic DNA from tissues, and subsequent hybridization with labelled cloned LAV provival DNA. Hybridization in situ can also be used.

Lymphatic fluids and tissues and other non-lymphatic tissues of humans, primates and other mammalian species can also be screened to see if other ~~evolutionnary~~ evolutionary related retrovirus exist. The methods referred to hereabove can be used, although hybridization and washings would be done under non-stringent conditions.

The DNA according to the invention can also be used ~~also~~ for achieving the expression of LAV viral antigens for diagnostic purposes.

The invention also relates to the polypeptides themselves which can be expressed by the different DNAs of the inventions, particularly by the ORFs or fragments thereof, in appropriate hosts, particularly procaryotic or eucaryotic hosts, after transformation thereof with a suitable vector previously modified by the corresponding DNAs.

These polypeptides can be used as diagnostic tools, particularly for the detection of antibodies in biological media, particularly in sera or tissues of persons afflicted with pre-AIDS or AIDS, or simply carrying antibodies in the absence of any apparent disorders. Conversely, the different peptides according to this invention can be used themselves for the production of antibodies, preferably monoclonal antibodies specific of the different peptides respectively. For the production of hybridomas secreting said monoclonal antibodies, conventional production and screening methods are used. These monoclonal antibodies, which themselves are part of

the invention, then provide very useful tools for the identification and even determination of relative proportions of the different polypeptides or proteins in biological samples, particularly human samples containing LAV or related viruses.

Thus, all of the above peptides can be used in diagnostics as sources of immunogens or antigens free of viral particles, produced using non-permissive systems, and thus of little or no biohazard risk.

The invention further relates to the hosts (procaryotic or eucaryotic cells) which are transformed by the ~~above mentioned~~ above-mentioned recombinants and which are capable of expressing said DNA fragments.

Finally, it also relates to vaccine compositions whose active principle is to be constituted by any of the expressed antigens, i.e. whole antigens, fusion polypeptides or oligopeptides, in association with a suitable pharmaceutical or physiologically acceptable carrier.

Preferably, the active principles to be considered in that field consist of the peptides containing less than 250 ~~aminoacid~~ amino acid units, preferably less than 150 as deducible ~~for~~ from the complete ~~genomas~~ genomes of LAV, and even more preferably those peptides which contain one or more groups selected from N-X-S and N-X-T as defined above. Preferred peptides for use in the production of vaccinating principles are peptides (a) to (f) as defined above. By way of example having no limitative character, there may be mentioned that suitable dosages of the vaccine compositions are those which enable administration to the host, particularly human host ranging from 10 to 500 micrograms per kg, for instance 50 to 100 micrograms per kg.

For the purpose of clarity, figs. 19 to 26 are added. ~~reference~~ Reference may be made thereto in case of difficulties of reading blurred parts of figs. 4 to 12.

Needless to say that figs. 19-26 are merely a reiteration of the whole DNA sequence of the LAV genome ~~genoma~~.

Finally, the invention also concerns vectors for the transformation of eucaryotic cells of human origin, particularly lymphocytes, the polymerases of which are capable of recognizing the LTRs of LAV. Particularly, said vectors are characterized by the presence of a LAV LTR therein, said LTR being then active as a promoter enabling the efficient transcription and translation in a suitable host of the above defined ~~of~~ DNA insert coding for a determined protein placed under its controls.

Needless to say that the invention extends to all variants of genomes and corresponding DNA fragments (ORFs) having substantially equivalent properties, all of said genomes belonging to retroviruses which can be considered as equivalents of LAV.